(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 February 2003 (27.02.2003)

PCT

(10) International Publication Number
WO 03/017680 A1

(51) International Patent Classification[7]: H04N 13/00, 7/14

(21) International Application Number: PCT/IB02/02961

(22) International Filing Date: 9 July 2002 (09.07.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
01203108.4 15 August 2001 (15.08.2001) EP

(71) Applicant: KONINKLIJKE PHILIPS ELECTRONICS N.V. [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).

(72) Inventors: VAN GEEST, Bartolomeus, W., D.; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). OP DE BEECK, Marc, J., R.; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).

(74) Agent: GROENENDAAL, Antonius, W., M.; Internationaal Octrooibureau B.V., Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).

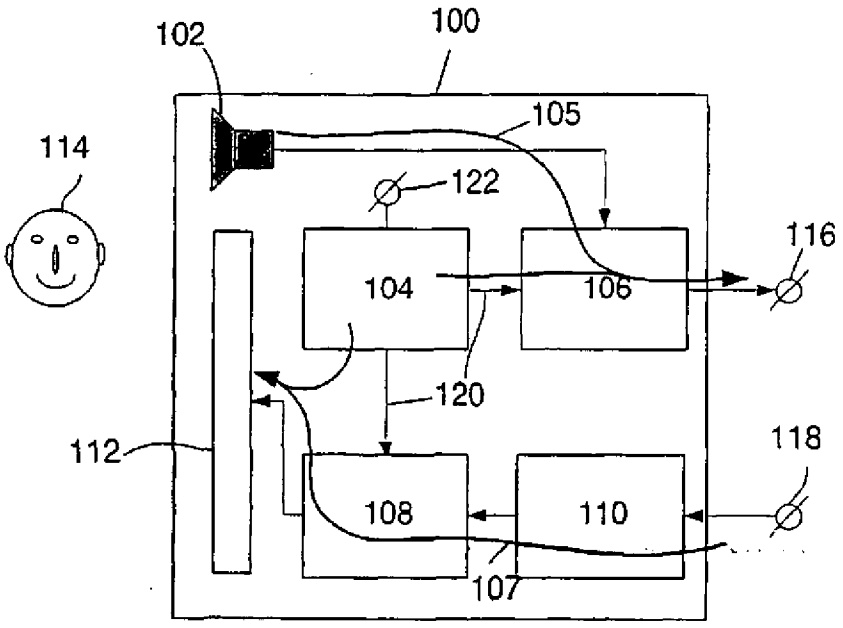
(81) Designated States *(national)*: CN, JP, KR.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: 3D VIDEO CONFERENCING SYSTEM



(57) Abstract: A 3D video conferencing station (100) comprises a video camera (102) for capturing video signals and a depth map calculator (104) for creating a depth map (120) of a user (114) of the 3D video conferencing station (100). Thevideo signals together with the depth map are transmitted as 3D video data. The 3D video conferencing station (100) further comprises a stereoscopic display device (112) for displaying stereo images, which are calculated on the basis of received 3D video data. The depth map which is generated by the depth map calculator (104) is also used for estimating the position of the user (114) in order to control the calculation of the stereo images.

WO 03/017680 A1

3D VIDEO CONFERENCING SYSTEM

The invention relates to a 3D video conferencing station and method.

In US 5,872,590 an image display apparatus is described that allows stereoscopic video images to be observed. A position of an observer in a stereoscopic observing region is detected by a position detecting unit. A right-eye image and a left-eye image are formed by an image forming unit and displayed on a display. By setting an aperture position of a projection optical system, the right-eye image is projected to the right-eye position of the observer and the left-eye image is projected to the left-eye position, thereby allowing a stereoscopic image to be observed. The position detecting unit is either based on two magnetic sensors or two ultrasonic sensors. The disadvantage of the approach based on the magnetic sensors is that it is invasive: a special tag is attached to the observer. The disadvantage of the position detecting unit based on ultrasonic sensors is that it is less robust.

It is an object of the invention to provide a 3D video conferencing station and method that are designed to track the position of a user relatively easy. To this end, the invention provides a 3D video conferencing station and method as defined by the independent claims. Another aspect of the invention provides an eye-tracking method and device that are especially useful for such a 3D video conferencing station. The dependent claims define advantageous embodiments.

The object of the invention is achieved in that the 3D video conferencing station comprises:

- a video camera for capturing video signals;
- depth means for creating a depth map of a user of the 3D video conferencing station;
- a transmitter for transmitting 3D video data which is based on the video signals and the depth map; and

- a stereoscopic display device for displaying stereo images, which are calculated on the basis of the depth map and received video signals.

The main advantage of the 3D video conferencing station according to the invention is that the depth map is used for two different tasks of the 3D video conferencing station:

- The first task is to create 3D video data based on the video signals captured by the camera. The 3D video data enables 3D visualization of the user on a second 3D video conferencing station to which the 3D video data is transmitted. In other words, the depth map is an essential component of the 3D video data which is used by the second 3D video conferencing station. The second 3D video conferencing station will typically be situated on another location.

- The second task is to track the position of the user relative to the stereoscopic display device. This position is required to calculate the appropriate images pairs, i.e. stereo images. Based on the position of the user the images pairs are calculated and rendered from 3D video data which is received from the second 3D video conferencing station. In other words, the depth map is used to track the position of the user, or more precisely his eyes, in order to control a part of the 3D video conferencing station itself. Based on the depth map it is relatively easy to determine the position of the user. Additional equipment, e.g. sensors, for user or eye-tracking are not required in the 3D video conferencing station according to the invention.

In an embodiment of the 3D video conferencing station according to the invention, the depth means are arranged to receive a signal from a depth sensor which is registered with the camera. They are both geometrically and in time space linked. The advantage of using a depth sensor which is registered with the camera is that relatively easy depth maps are captured with a relatively high quality.

In another embodiment of the 3D video conferencing station according to the invention, the depth means are designed to create the depth map on the basis of the video signals. The advantage of using the video signals to create the depth map is that no additional depth sensor is required.

An embodiment of the 3D video conferencing station according to the invention comprises a processor for detecting a position of a nose of the user of the 3D video conferencing station by analyzing the depth map. In order to control the creation of stereo pairs which are displayed on the stereoscopic display device, it is important to know the position of the eyes of the user as good as possible. The position of the nose of the user is a

good indication of the position of the eyes. The nose can be found relatively easily in a depth map of the user.

In an embodiment of the 3D video conferencing station according to the invention the processor is designed to detect the position of the nose of the user of the 3D video conferencing station by searching for a maximum or a minimum depth value of the depth map. The nose of the user will normally be the part of the face which is located closest to the camera. Hence it corresponds with a maximum or minimum depth value of the depth map, depending on the coordinate system. Finding a maximum or minimum depth value of the depth map is a relatively simple operation.

In another embodiment of the 3D video conferencing station according to the invention, the processor is designed to detect the position of the nose of the user of the 3D video conferencing station by comparing depth values of the depth map with a model of a human face. In the case that the head of the user is at a tilt relative to the camera it might be that the forehead or the chin of the user has a lower/ higher depth value or than the depth value corresponding with the nose. By taking multiple depth values of the depth map into account and match these with a model of a human face, a more robust nose detection is achieved.

In another embodiment of the 3D video conferencing station according to the invention, the processor is designed to detect an orientation of a head of the user of the 3D video conferencing station by calculating an average derivative of depth values of a region of the depth map that corresponds with a part of a face of the user. In the case that the head of the user is turned relative to the camera the distance between the position of the nose and the left eye can deviate relatively much from the distance between the position of the nose and the right eye. In that case the position of the nose is a less good indication of the positions of the individual eyes. With the average derivative of depth values of a region of the depth map that corresponds with a part of the face of the user, the orientation of the head can be estimated. With information of the orientation of the head and the position of the nose the position of the eyes can be estimated more accurately.

In an embodiment of the 3D video conferencing station according to the invention, the processor is designed to detect a first position of a left eye and a second position of a right eye based on the position of the nose of the user of the 3D video conferencing station. Information of the position of the nose of the users is a good starting point for controlling the creation of images pairs. With knowledge of the actual position of the eyes an improved control is achieved.

In a preferred embodiment of the 3D video conferencing station according to the invention, the processor is designed to detect the first position of the left eye and the second position of the right eye based on the video signals. Besides the depth map, also the video signals are input for this embodiment of the processor of the 3D video conferencing station. The luminance and optionally chrominance values of the pixels corresponding to the video signals provide additional data, which is very useful to improve the robustness of the eye detection.

In the article "Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D models" of M. La Cascia, et al. in IEEE Transactions on pattern analysis and machine intelligence, Vol. 22, No. 4, April 2000 a technique for 3D head tracking under varying illumination conditions is described. The head is modeled as a texture mapped cylinder. Tracking is formulated as an image registration problem in the cylinders texture mapped image. The resulting dynamic texture map provides a stabilized view of the face that can be used for eye-tracking. The method described in this article is relatively complex for eye tracking compared with the method as performed by the 3D video conferencing station according to the invention.

Modifications of the video conferencing station and variations thereof may correspond to modifications and variations thereof of the eye-tracker and of the method of eye-tracking described.

These and other aspects of the 3D video conferencing station and method according to the invention will become apparent from and will be elucidated with respect to the implementations and embodiments described hereinafter and with reference to the accompanying drawings, wherein:

Fig. 1A schematically shows an embodiment of the 3D video conferencing station;

Fig. 1B schematically shows an embodiment of the 3D video conferencing station comprising a depth sensor;

Fig. 1C schematically shows an embodiment of the 3D video conferencing station which is designed to calculate a depth map on the basis of video signals;

Fig. 2A schematically shows an embodiment of the 3D video conferencing station comprising a processor which is designed to detect a position of a nose of the user;

Fig. 2B schematically shows an embodiment of the 3D video conferencing station comprising a processor which is designed to detect the position of the eyes of the user; and

Fig. 3 schematically shows a 3D video conferencing system comprising two 3D video conferencing stations according to the invention.

Corresponding reference numerals have the same meaning in all of the Figs..

Fig. 1A schematically shows an embodiment of the 3D video conferencing station 100 comprising:

- a video camera 102 for capturing video signals;
- a depth map calculator 104 for creating a depth map 120 of a user 114 of the 3D video conferencing station 100;
- a transmitter 106 for transmitting 3D video data which is based on the video signals and the depth map 120;
- a receiver 110 for receiving 3D video data which has been acquired by a second 3D video conferencing station 301; and
- a stereo images generator 108 for generating stereo images based on the 3D video data which is received by the receiver 110. The stereo images generator 108 requires information about the position of the user 114. This information is retrieved from the depth map 120 which is generated by the depth map calculator 104 ; and
- a stereoscopic display device 112 for displaying the stereo images, which are generated by the stereo images generator 108.

Two main data flows can be distinguished in the 3D video conferencing station 100:

- Outgoing data flow 105: First there are video signals which are captured by the video camera 102. These video signals are enhanced with depth maps, resulting in 3D video data. The depth maps are generated by the depth map calculator 104. This 3D video data is transmitted by the transmitter 106. This 3D video data is provided by the 3D video conferencing station 100 at its output connector 116. Optionally the 3D video is encoded, e.g. according to one of the MPEG standard formats.
- Incoming data flow 107: Second there is 3D video data which is generated by the second 3D video conferencing station 301. This 3D video data is provided at the input connector 118 and received by the receiver 110. The stereo images generator 108 renders

stereo images based on this 3D video data, in dependence of the position of the user 114. The position of the user is determined based on the depth map 120. The generated stereo images are displayed by the stereoscopic display device 112.

Fig. 1B schematically shows an embodiment of the 3D video conferencing station 101 comprising a depth sensor 124. The signals of the depth sensor 124 are provided to the input connector 122 of the depth map calculator 104 for creating a depth map. In this case the depth sensor provides signals which are related to propagation times of waves, e.g. ultrasonic or infrared, which are respectively generated by the depth sensor, reflected by the user and received by the depth sensor. The major task of the depth map calculator 104 is conversion of the signals which are related to propagation times to depth values. Other tasks are e.g. synchronization and temporarily storage of data. Synchronization is required with the video signals which are generated by the video camera 102. The depth sensor 124 and the video camera 102 are both geometrically and in time space linked. In other words, pixels of the video images corresponding with the video signals which are generated by the video camera 102 are spatially and temporarily correlated with the depth maps as created by the depth map calculator 104. Notice that there are systems commercially available which combine the functions of the video camera 102, the depth sensor 124 and the depth map calculator 104, e.g. ZCam™ from 3DV Systems.

Fig. 1C schematically shows an embodiment of the 3D video conferencing station 103 which is designed to calculate a depth map 120 on the basis of video signals. In this case the video signals which are captured by the video camera 102 are also provided to the depth map calculator 104. By applying geometrical relations, the depth information can be deduced from motion. This concept is e.g. described by P. Wilinski and K. van Overveld in the article "Depth from motion using confidence based block matching" in Proceedings of Image and Multidimensional Signal Processing Workshop, pages 159-162, Alpbach, Austria, 1998, and in WO99/40726. All apparent motion in the series of images results from parallax. Differences in motion between one segment and another indicate a depth difference. Analyzing two consecutive images, the parallax between a given image segment at time t and the same segment at t+1 can be computed. This parallax corresponds to the motion of different parts of the scene. In the case of translation of the camera, objects in the foreground move more than those in the background. It is important that there is movement of the user relative to the camera. Whether the camera moves or the user is in principle irrelevant. Optionally multiple cameras are used to capture the video signals. The method of creating a depth map is conceptually the same in that case. Estimating the depth map which is used in

the 3D video conferencing station 103 is not limited to the method as described in the cited article, but other depth estimation methods can also be used.

Fig. 2A schematically shows an embodiment of the 3D video conferencing station 200 comprising a processor 202 which is designed to detect a position of the nose of the user. The processor 202 requires a depth map 120 as input and provides coordinates 202 of the position of the nose to the stereo images generator 108. The coordinate system is defined such that points which are close to the 3D video conferencing system 200 have a low depth value, i.e. z-coordinate. The user is looking at the stereoscopic display device 112. The video camera 102 and/or the depth sensor 124 are aligned with the stereoscopic display device 112. As a consequence the nose of the user has a relatively low z-coordinate. The working of the processor is as follows. Each predetermined time interval a new depth map is processed. In the each depth map the lowest depth value is searched. The corresponding x and y-coordinates of the tip of the nose are automatically known then too.

Optionally a segmentation is performed of the depth map to determine a region of interest in the depth map corresponding to the face of the user. This segmentation is e.g. performed by means of a threshold operation, i.e. only low depth values are remained. It is assumed that relatively high depth values correspond with other objects in the scene in which the user is located, e.g. the background. The depth values of the region of interest are compared with a model of a human face. In that case the coordinates of the nose are searched with a template matching technique.

Optionally the orientation of the head of the user is estimated. This can be achieved by calculating derivatives of depth values of the region of interest. The assumption is that the head of the user is relatively symmetrical. Comparing derivatives of a left portion of the region of interest with derivatives of a right portion of the region enables to estimate the orientation of the head.

Optionally the location of the nose in depth map N is determined by making use of the detected position based on depth map N-1 which was acquired before. The advantages of this approach are that the detection of the nose of the user can be faster and more robust.

Fig. 2B schematically shows an embodiment of the 3D video conferencing station 201 comprising a processor 202 which is designed to detect the position of the eyes of the user. The processor 202 requires a depth map 120 as input and provides coordinates 204 of the position of the eyes to the stereo images generator 108. With respect to the positions of the right and left eye of the user, since the interval between the human eyes statistically lies

in a range from 32.5 mm to 97.5 mm, an interval W between the two eyes is set to, for example W=60 mm and it is sufficient to obtain the x-coordinate values of the respective eyes by adding or subtracting W/2 to/from the coordinate value at the center position between the both eyes. This center position can correspond to the x-coordinate of the tip of the nose. Optionally this center position is based on the x-coordinate of the tip of the nose, but by taking into account the orientation of the head. In that case the distances, in projection, from the eyes to the tip of the nose are not mutually equal.

Optionally the video signals are also input for the processor. The luminance and chrominance values of the pixels corresponding to the video signals provide additional data, which is very useful to improve the robustness of the eye detection. Typically eyes result in high contrast in the images corresponding to the video signals. Also the color of the eyes deviates relatively much from the color of the skin.

Fig. 3 schematically shows a 3D video conferencing system 300 comprising two 3D video conferencing stations 100 and 301 according to the invention. The working of the 3D video conferencing stations 100 and 301 is described in connection with one of the Figs. 1A, 1B, 2A or 2B. The 3D video conferencing stations 100 and 301 can be connected by means of a private communication link. A public communication link, e.g. internet is also possible. Preferably the 3D video conferencing system allows for concurrent communication between the 3D video conferencing stations 100 and 301. Optionally the 3D video conferencing system 100 comprises more 3D video conferencing stations than the two 3D video conferencing stations 100 and 301.

With the depth map calculator 104 and the processor 202 as shown in Fig. 2A and 2B an eye-tracker can be constructed which can also be used in various types of systems, e.g. comprising a stereoscopic display device.

Stereoscopic video is used as example of 3D video in the embodiments. Other 3D visualizations are also possible. E.g. a standard 2D display on which a rendered 3D model is rotated in dependence of the observer. Alternatively a multiview display, e.g. Philips 3D-LCD in combination with multiple video channels can be used. These multiple views are generally projected in fixed directions. Knowledge of the position of the observer can be applied to control these directions. A third alternative are multi-depth-layer displays.

It should be noted that the above-mentioned embodiments illustrate rather than limit the invention and that those skilled in the art will be able to design alternative embodiments without departing from the scope of the appended claims. In the claims, any reference signs placed between parentheses shall not be constructed as limiting the claim.

The word 'comprising' does not exclude the presence of elements or steps not listed in a claim. The word "a" or "an" preceding an element does not exclude the presence of a plurality of such elements. The invention can be implemented by means of hardware comprising several distinct elements and by means of a suitable programmed computer. In the unit claims enumerating several means, several of these means can be embodied by one and the same item of hardware.

CLAIMS:

1. A 3D video conferencing station (100) comprising:

- a video camera (102) for capturing video signals;

- depth means (104) for creating a depth map (120) of a user (114) of the 3D video conferencing station (100);

- a transmitter (106) for transmitting 3D video data which is based on the video signals and the depth map (120); and

- a stereoscopic display device (112) for displaying stereo images, which are calculated on the basis of the depth map (120) and received video signals (110).

2. A 3D video conferencing station (101) as claimed in Claim 1, characterized in that the depth means (104) are arranged to receive a signal from a depth sensor (124) which is registered with the camera.

3. A 3D video conferencing station (103) as claimed in Claim 1, characterized in that the depth means (104) are designed to create the depth map (120) on the basis of the video signals.

4. A 3D video conferencing station (200) as claimed in Claim 1, characterized in that the 3D video conferencing station (200) comprises a processor (202) for detecting a position of a nose of the user (114) of the 3D video conferencing station (200) by analyzing the depth map (120).

5. A 3D video conferencing station (200) as claimed in Claim 4, characterized in that the processor (202) is designed to detect the position of the nose of the user (114) of the 3D video conferencing station (100) by searching for a maximum or a minimum depth value of the depth map (120).

6. A 3D video conferencing station (200) as claimed in Claim 4, characterized in that the processor is designed to detect the position of the nose of the user (114) of the 3D

video conferencing station (100) by comparing depth values of the depth map (120) with a model of a human face.

7. A 3D video conferencing station (200) as claimed in Claim 4, characterized in that the processor (202) is designed to detect an orientation of a head of the user (114) of the 3D video conferencing station (100) by calculating an average derivative of depth values of a region of the depth map (120) that corresponds with a part of a face of the user (114).

8. A 3D video conferencing station (200) as claimed in Claim 4, characterized in that the processor (202) is designed to detect a first position of a left eye and a second position of a right eye based on the position of the nose of the user (114) of the 3D video conferencing station (200).

9. A 3D video conferencing station (100) as claimed in Claim 8, characterized in that the processor (202) is designed to detect the first position of the left eye and the second position of the right eye based on the video signals.

10. A 3D video conferencing method (100) comprising:

- capturing (102) video signals;
- creating (104) a depth map (120) of a user (114) of the 3D video conferencing station (100);
- transmitting (106) 3D video data which is based on the video signals and the depth map (120); and
- displaying (112) stereo images, which are calculated on the basis of the depth map (120) and received video signals (110).

11. An eye-tracker (104,202) for estimating a first position of a left eye and a second position of a right eye, characterized in that the eye tracker (104,202) is designed to estimate the first position of the left eye and the second position of the right eye on the basis of a depth map (120).

12. A method of eye-tracking to estimate a first position of a left eye and a second position of a right eye, characterized in that the first position of the left eye and the second position of the right eye are estimated on the basis of a depth map (120).

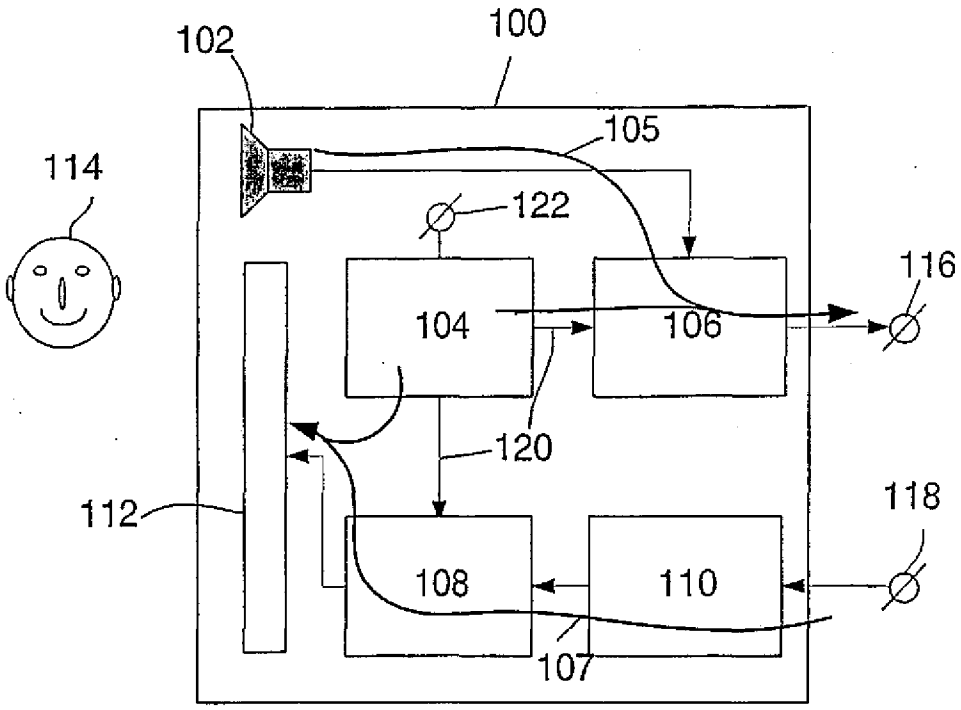

102
100
105
114
122
116
104
106
120
112
118
108
110
107

FIG. 1A

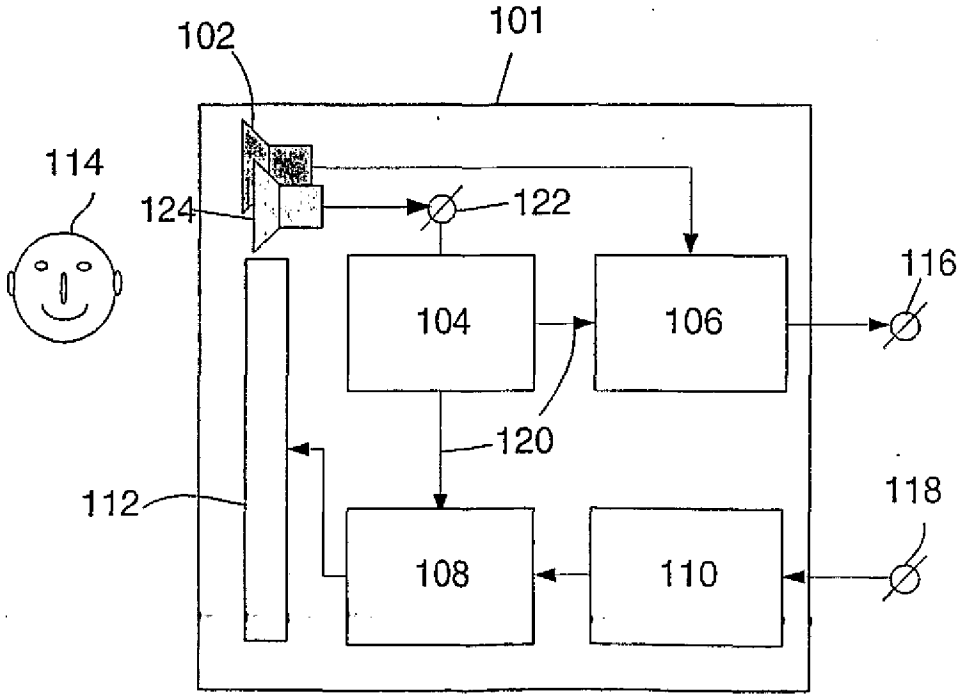

102
101
114
124
122
116
104
106
120
112
118
108
110

FIG. 1B

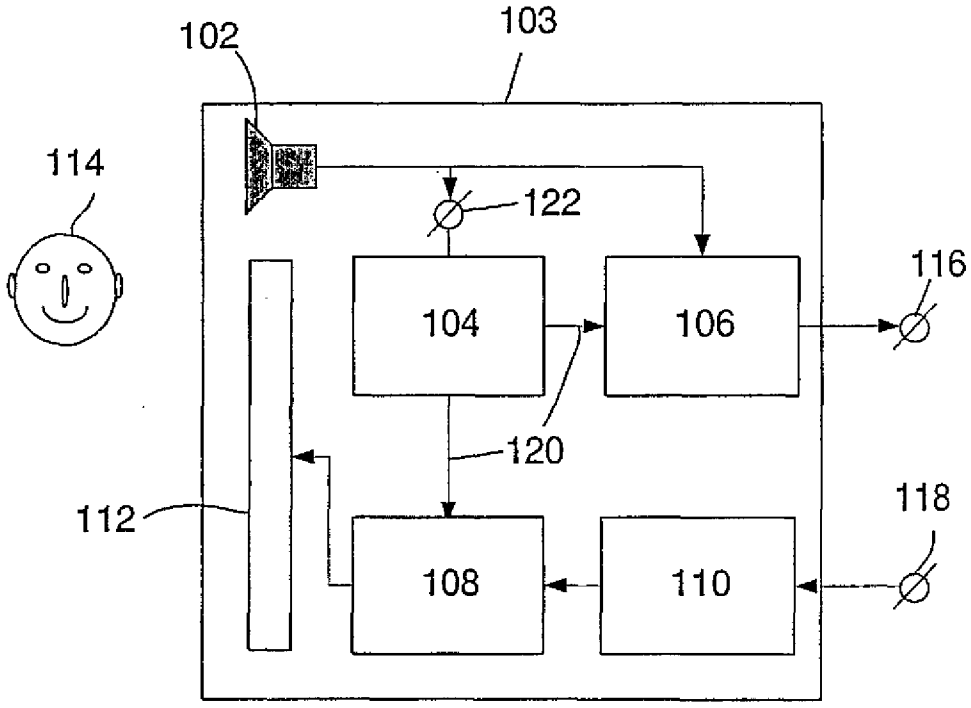


FIG. 1C

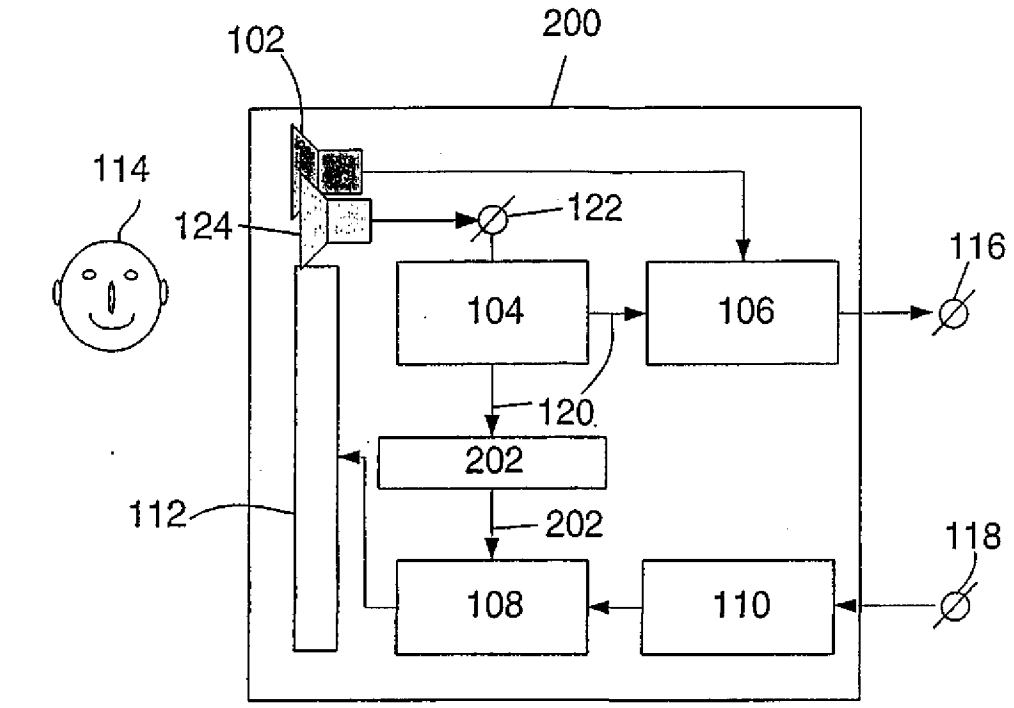

102
200
114
124
122
116
104
106
120
202
112
202
118
108
110

FIG. 2A

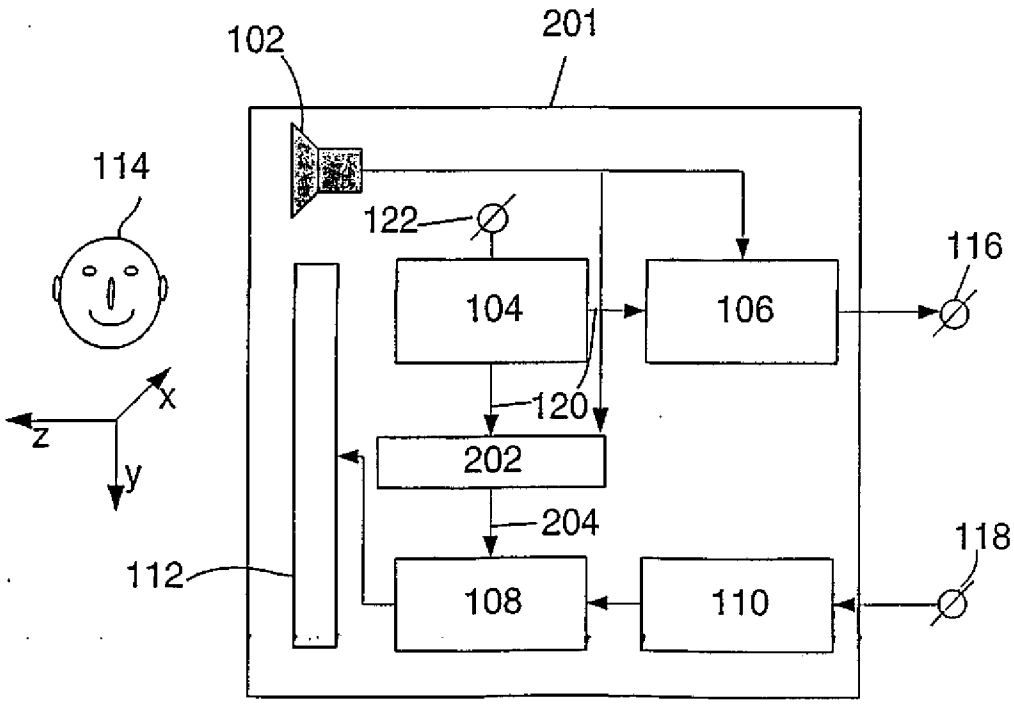

102
201
114
122
116
104
106
120
x
z
y
202
204
112
118
108
110

FIG. 2B

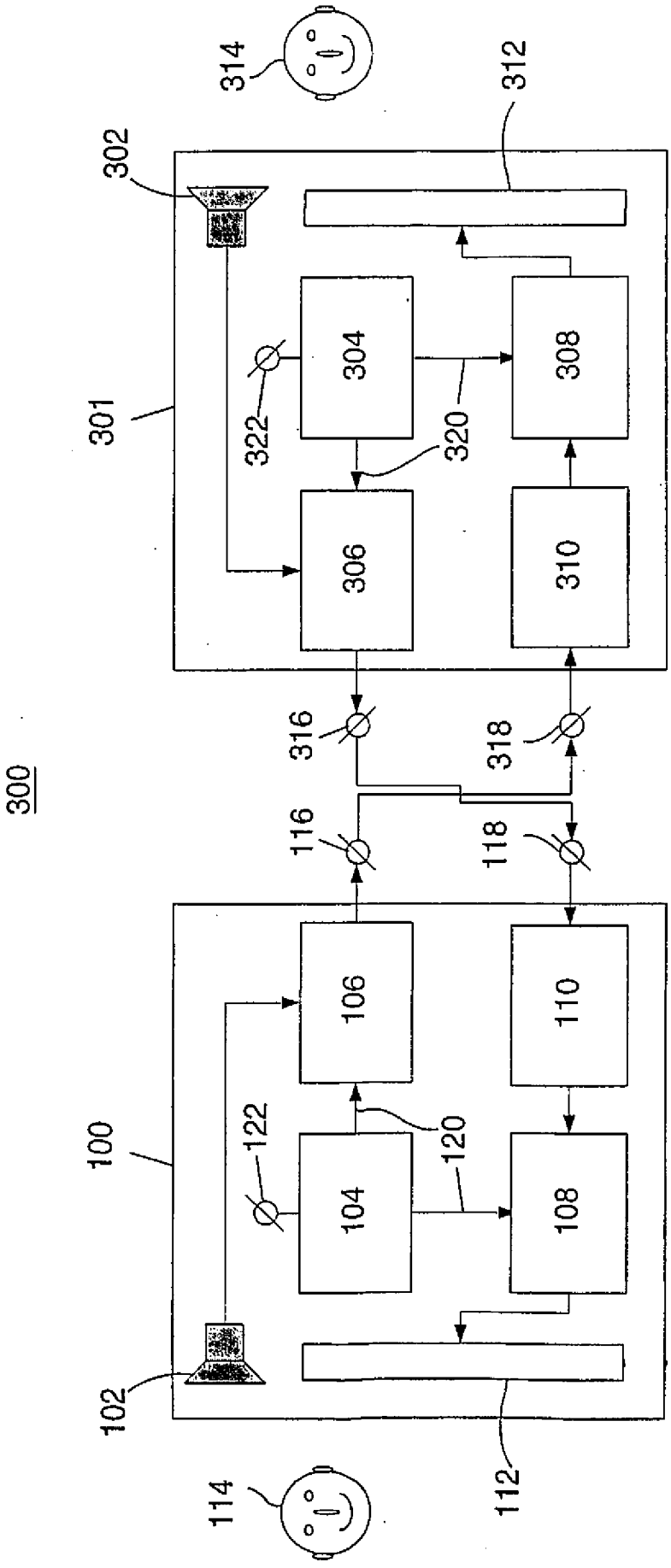

300
100
102
104
106
108
110
112
114
116
118
120
122
301
302
304
306
308
310
312
314
316
318
320
322

FIG. 3

INTERNATIONAL SEARCH REPORT

tional Application No
PCT/IB 02/02961

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 H04N13/00 H04N7/14

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 H04N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ, INSPEC, COMPENDEX

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 6 055 012 A (HASKELL BARIN GEOFFRY ET AL) 25 April 2000 (2000-04-25)	1-3,10
Y	column 6, line 11 -column 7, line 14; figure 1 column 10, line 62 -column 13, line 6; figures 15,16	4,6-9

	-/--	

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 October 2002	22/10/2002

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	De Paepe, W

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
X	TALMI K ET AL: "Eye and gaze tracking for visually controlled interactive stereoscopic displays - an experiment study on the subjective effects of disparity magnitude and depth of focus" SIGNAL PROCESSING. IMAGE COMMUNICATION, ELSEVIER SCIENCE PUBLISHERS, AMSTERDAM, NL, vol. 14, no. 10, August 1999 (1999-08), pages 799-810, XP004173766 ISSN: 0923-5965 2. System overview 3. Eye tracker	11,12
Y	WO 99 57900 A (MYERS JOHN KARL) 11 November 1999 (1999-11-11) page 52, line 17 -page 53, line 2; figure 14 page 20, line 33 -page 21, line 33	4,6-9
A	WO 01 29767 A (KONINKL PHILIPS ELECTRONICS NV) 26 April 2001 (2001-04-26)	

1

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 6055012	A	25-04-2000	NONE		
WO 9957900	A	11-11-1999	AU	4307499 A	23-11-1999
			WO	9957900 A1	11-11-1999
WO 0129767	A	26-04-2001	WO	0129767 A2	26-04-2001
			EP	1190385 A2	27-03-2002